International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Special Issue: 1 Volume No.7 April – 2023
DOI:10.46647/ijetms.2023.v07si01.076 ISSN: 2581-4621

Emotion Predictor Using Sentiment Analysis

S Venkata Lakshm1[1] , A Rajesh[2], K Laya[3], P Bala Chandar[4], E Muni Teja[5]
Assistant professor, student, Department of CSE, AITS, Tirupati
sibbalalakshmiaits@gmail.com

Abstract
Now-a-days social networks are booming, so a huge amount of data is generated. Millions of people share their views on micro-blogging websites every day because they contain short and simple expressions. In this project, we will discuss a paradigm for extracting sentiments from the famous micro-blogging service Twitter, where users post their opinions on all kinds of topics. In this project, we will discuss the existing analysis of the Twitter dataset using data mining approaches, such as using sentiment analysis algorithms using machine learning algorithms. An approach is presented that automatically classifies the sentiments of tweets from the Twitter dataset. These messages or tweets are classified as positive, negative or neutral with respect to a query term.
Keywords: Microblogging, Twitter, sentiment, classifiers, sentiment analysis.

Introduction

We know that there are nearly 111 micro-blogging spots. Micro-blogging spots are nothing but social media spots where druggies make short and frequent posts. Twitter is one of the most popular micro-blogging services where druggies can read and post dispatches with a length of 148 characters. Twitter dispatches are also appertained to as tweets. We'll use these tweets as raw data. We'll use a system that automatically classifies tweets into positive, negative, or neutral sentiments. With the help of sentiment analysis, the client can know the feedback about a product or service before making a purchase. The company can use sentiment analysis to know the guests' opinion about its products so that it can dissect the client satisfaction and ameliorate its product consequently. Sentiment analysis has come a popular exploration area in computational linguistics due to the explosion of sentiment information from social websites (e.g., Twitter and Facebook), online forums, and blogs. In this design, we use colorful machine learning classifiers similar as Naive Bayes, KNN Classifier and Support Vector Machines(SVM) are the machine learning classifiers. Sentiment analysis is veritably sphere specific; the operation used for twitter cannot be used for Facebook. For Twitter, this is particularly problematic.

For illustration," The food was great, but the service was terrible." In this case, the computer is confused with the result of the sentiment.

Related Work

1. Algorithms:

There are three different machine learning algorithms who achieved great success for text categorization which are as follows:

1.1 Multinomial Naive Bayes:

This is substantially used for document bracket problem, i.e. whether a document belongs to the order of sports, politics, technology etc. The features predictors used by the classifier are the frequency of the words present in the document.

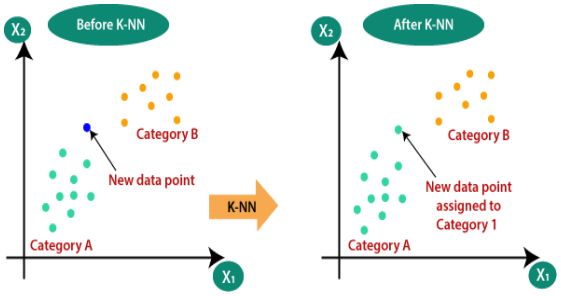
1.2 K-Nearest Neighbor Algorithm:

The k- nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised literacy classifier, which uses propinquity to make groups or prognostications about the grouping of an individual data point. While it can be used for either regression or bracket problems, it's generally used as a bracket algorithm, working off the supposition that analogous points can be set up near one another.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Special Issue: 1 Volume No.7 April – 2023
DOI:10.46647/ijetms.2023.v07si01.076 ISSN: 2581-4621

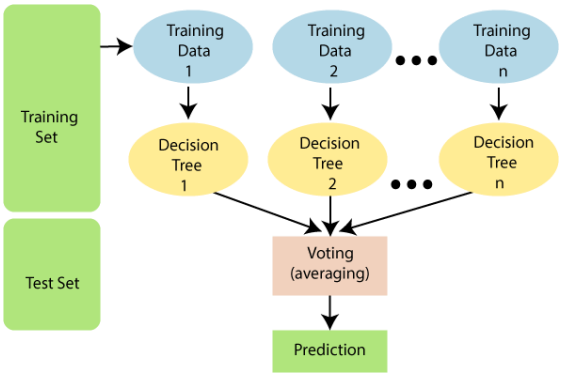
Why do we need a K-NN Algorithm?

Suppose there are two orders, i.e., order A and order B, and we have a new data point x1, so this data point will lie in which of these orders. To break this type of problem, we need a K- NN algorithm. With the help of K- NN, we can easily identify the order or class of a particular dataset. Consider the below illustration



1.3 Random Forest Algorithm:

Random Forest is a popular machine knowledge algorithm that belongs to the supervised knowledge fashion. It can be used for both Bracket and Retrogression problems in ML. It's predicated on the generality of ensemble knowledge, which is a process of combining multiple classifiers to break a complex problem and to meliorate the performance of the model. As the name suggests," Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to meliorate the predictive delicacy of that dataset." rather of counting on one decision tree, the arbitrary timber takes the prophecy from each tree and predicated on the maturity votes of prognostications, and it predicts the final affair.



2. Proposed System:

The following way will explain the process of the proposed system.

1. Retrieval of tweets
2. Preprocessing of the uprooted data
3. Resemblant processing
4. Emotional State Affair
5. Delicacy of the algorithm

2.1 Retrieval of tweets:

Since Twitter is the largest part of the social networking website, it consists of colorful blogs related to different motifs around the world. Rather than taking whole blogs, we prefer to search for a specific content and download all the web runners and also prize them in the form of textbook lines using the mining tool Weka, which provides a sentiment classifier.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Special Issue: 1 Volume No.7 April – 2023
DOI:10.46647/ijetms.2023.v07si01.076 ISSN: 2581-4621

2.2 Preprocessing of the uprooted data:
After reacquiring the tweets, the sentiment analysis tool is applied to the raw tweets, but this results in veritably poor performance in utmost cases. Thus, preprocessing ways are necessary to achieve better results. We prize tweets, i.e., short dispatches from Twitter, which are used as raw data. This raw data needs to be preprocessed. Therefore, preprocessing includes the following way that construct n- grams

i) Filtering: Filtering is nothing other than drawing the raw data. In this step, URL links (e.g., http//twitter.com), special words in Twitter(e.g.," RT" meaning ReTweet), usernames in Twitter(e.g.,@Ron-@ symbol indicating a username), and emoticons are removed.

ii) Tokenization: Tokenization is nothing but the segmentation of rulings. In this step, textbook is tokenized using spaces and punctuation marks to form a vessel of words.

iii) Junking of stop words: Articles like" a"," one"," the" and other stop words like" to"," of"," is"," are"," this"," for" are removed in this step.

iv) Construction of n- grams: A set of n- grams can be constructed from successive words. Negation words like" no"," not" are attached to a word that follows or precedes it. For illustration," I don't like remix music" has two bigrams" I like not"," like not"," like not remix music". So, the delicacy of bracket improves by such a procedure, because negation plays an important part in sentiment analysis. The paper presents that negation must be considered because it's a veritably common verbal construction that affects opposition.

2.3 Resemblant processing:
The sentiment classifier that classifies sentiments creates a multinomial Naïve Bayes classifier or a Random Forest. Training the bracket data is the main reason for this step. Every database contains retired information that can be used for decision timber. Bracket and vaticination are two forms of data analysis that can be used to prize models that describe important data and unborn trends. Bracket is about chancing a set of models or functions that describe and distinguish data classes or generalities in order to use the model to prognosticate the class of objects whose class marker is unknown.

2.4 Emotional state affair:
The proposed system interprets whether the tweet is positive, negative, or neutral grounded on the points assigned in the wordbook.

2.5 Delicacy of the algorithm:
Using the confusion matrix and the bracket report, we can determine the delicacy of the machine learning algorithms.

3. Result:
We can find the delicacy of each algorithm by chancing confusion matrix and classification report.

A. Confusion Matrix:
In the field of machine literacy and specifically the problem of statistical bracket, a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, generally a supervised literacy one (in unsupervised literacy it's generally called a matching matrix). Each row of the matrix represents the cases in an factual class while each column represents the cases in a prognosticated class, or vice versa – both variants are set up in the literature. The name stems from the fact that it makes it easy to see whether the system is confusing two classes (i.e., generally mislabeling one as another). It's a special kind of contingency table, with two confines (" factual" and" prognosticated"), and identical sets of" classes" in both confines(each combination of dimension and class is a variable in the contingency table).

B. Classification Report: It displays your model's perfection, recall, F1 score and support. Overall performance of our trained model. To understand the classification report of a machine literacy model, you need to know all of the criteria displayed in the report.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Special Issue: 1 Volume No.7 April – 2023
DOI:10.46647/ijetms.2023.v07si01.076 ISSN: 2581-4621

By chancing confusion matrix and classification report we can find that

Metrics	Definition
Precision	Precision is defined as the ratio of true positives to the sum of true and false positives.
Recall	Recall is defined as the ratio of true positives to the sum of true positives and false negatives.
F1 Score	The F1 is the weighted harmonic mean of precision and recall. The closer the value of the F1 score is to 1.0, the better the expected performance of the model is.
Support	Support is the number of actual occurrences of the class in the dataset. It doesn't vary between models, it just diagnoses the performance evaluation process.

• The multinomial Naive Bayes algorithm yields an accuracy of 0.68
• KNN classifier yields an accuracy of 0.64
• Random Forest classifier gives an accuracy of 0.50
We can conclude that Multinomial Naive Bayes algorithm has high accuracy compared to KNN and Random Forest Classifier.

4. Conclusion:
Twitter is a sought-after micro-blogging service that lets you find out what's happening anytime, anywhere in the world. In the survey, we found that features related to social media can be used to predict Twitter sentiment.
We can also find out which machine learning algorithm has high accuracy and which algorithm works effectively for the provided dataset. Our
Proposed system infers the sentiments of tweets
that are taken from Twitter. The difficulty increases with the nuance and complexity of the opinions expressed. Product reviews, etc., are relatively straight forward. Books, movies, art and music are more difficult. We can also implement features such as emoticons, negation and the capitalization/ internationalization, as these have recently become an important part of the Internet.

5. References:
[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, "Sentiment Analysis of Twitter Data" Department of Computer Science, Columbia University, New York, 2009.
[2] Akshi Kumar and Teeja Mary Sebastian, "Sentiment Analysis on Twitter" department of Computer Engineering, Delhi Technological University, Delhi, India, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012.
[3] Luciano Barbosa and Junlan Feng, "Robust sentiment detection on twitter from biased and noisy data." Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44, 2010.
[4] Adam Bermingham and Alan Smeaton, "Classifying sentiment in microblogs: is brevity an advantage?" ACM, pages 1833–1836, 2010.
[5] R. Parikh and M. Movassate, "Sentiment Analysis of User- Generated Twitter Updates using Various Classification Techniques", CS224N Final Report, 2009

DOI:10.46647/ijetms.2023.v07si01.076 ISSN: 2581-4621

[6] Go, R. Bhayani, L.Huang. "Twitter Sentiment Classification Using Distant Supervision", Stanford University, Technical Paper, 2009
[7] Shai Shalev-Shwartz, Yoram Singer, Nathan, Srebro, Andrew Cotter "Pegasos: Primal Estimated subGrAdient SOlver for SVM", 2000.
[8] Chuan-Ju Wangz, Ming-Feng Tsaiy, Tse Liuy, Chin-Ting Changzy, "Financial Sentiment Analysis for Risk Prediction" Department of Computer Science & Program in Digital Content and Technology National Chengchi University Taipei 116, 2013.
[9] Changbo Wang, Zhao Xiao, Yuhua Liu, Yanru Xu, Aoying Zhou, and Kang Zhang, "SentiView: Sentiment Analysis and Visualization for Internet Popular Topics" IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, VOL. 43, NO. 6, NOVEMBER 2013.